Review on Retrieving Biological Sequence Alignment using Smith-Waterman Algorithm

Komal B. Deshmukh, M. U. Kharat

***Abstract*— Bioinformatics is one of the interdisciplinary research area. In various genome projects, huge biological sequences are available. Biological sequence analysis is fundamental operation in Bioinformatics and the goal is to find the similarity region based on comparing method is called as sequence alignment. In sequences alignment two methods are reported that are local alignment and global alignment. Smith-Waterman (SW) algorithm is providing optimal local alignment which has quadratic time and space complexity. In several real time applications huge sequences are aligned. In ordered to increase the efficiency, biological sequences are aligned over Graphical Processing Unit (GPU). Various systems applies parallel algorithm with use of GPU to accelerate application. CUDA Align algorithm is used to obtain full optimal alignment of biological sequences and the idea behind this is to obtain coordinate point of optimal alignment, iteratively perform same process to increase number of coordinate until it is sufficient to retrieve full optimal alignment.**

***Index Terms*— Bioinformatics, sequence alignment, Smith- waterman (SW) algorithm, GPU, CUDAlign.**

I. INTRODUCTION

Bioinformatics is one of integrative and important research area. It is interdisciplinary field includes computer science, biology, mathematics and statistics. It means standing with foot in the two worlds that is world of computer science and world of biological science. This is phenomenon of collecting and analyzing complex biological data that uses computer technology for management of biological data. The data is in terms of biological sequence. Computers are used to gather, store, integrate and analyze the large scale biological data and genetic information. The emphasis is on the use of computers because most of the tasks in genomic data analysis are highly repetitive or mathematically complex. Common activities in Bioinformatics include mapping and analyzing sequences, aligning sequences to compare them in order to discover function, structure and evolutionary information of particular new sequence.

Manuscript received January 07, 2015

Komal B. Deshmukh, Computer Department, MET BKC, Savitribai Phule Pune University, Nasik , India.

M. U. Kharat, Computer Department, MET BKC, Savitribai Phule Pune University, Nasik , India.

Biological sequence analysis is important operation in bioinformatics the goal is to find the similarity region based on comparing method is called as sequence alignment. If there is any new sequence is discovered, then that biological sequence is compared with other sequences available in genome database. So sequence comparison is one of the key operation is considered. The outcome of sequence comparison is measured in terms of score and alignment. Score is nothing but the similarity between the sequences. That is it provides one integer number indicate that how much two sequences are similar and alignment highlights the similarity between two sequences.

In various real time application alignments are reported in two ways that are local and global alignment. A global method is aiming from end-to-end of the sequences and there are two methods, functional as global method; Dot Plot and Needleman-Wunsch (NW) algorithm meanwhile, another alignment also carried two methods which are local alignment method. The methods are known as an exact method like Smith-Waterman algorithm and heuristic based approximate method like FASTA and BLAST . In local alignment method, both methods are attempted to identify the most similar region between pair or more sequences.

Smith-Waterman algorithm is providing optimal local alignment which has quadratic time and space complexity. This algorithm uses the dynamic programming approach. Using this approach it creates the dynamic programming matrices called DP matrix. There are some restriction related to SW algorithm. One of the important restrictions related to SW algorithm is that it requires quadratic space to store DP matrices. And another restrictive property is it requires quadratic time complexity. So far many efforts have been taken to reduce the time and space complexity of SW algorithm.

In ordered to increase the efficiency, biological sequences are aligned over Graphical processing Unit(GPU). Various systems applies parallel algorithm that uses GPU.

A. Sequence Analysis – Score and Alignment

Sequence Comparison is core function of Bioinformatics analysis. When the two sequences are compared, there is need to determine score and alignment between them. Figure 1 shows how to align two sequences.

```
A - C A T A C A
| | | | | | | |
A G C A G C - A
```

Fig 1: Sequence Alignment.

Score is measure of similarity between two sequences. Consider two sequences as seq1 and seq2, while calculating score need to consider following values for each instance.

1. val= +1, if both character of sequence are match (ma= +1)
2. val= -1, if both character of sequence are not match. (mi= -1)
3. val= -2 , if one of the character is space (G=-2)

Total score between the two sequences are calculated by summing up all val of each instance. Another measure of sequence analysis is Alignment. It basically highlights the similarity between two sequences. Sequences can be aligned locally or globally.

- Global Alignment - In this method two sequences are assume to be of same length. Two sequences are aligned from beginning to the end of both the sequences to find the best possible alignment. This approach provides very precise solution by considering whole length of sequence. But in some cases global alignment technique is insufficient because there is need to compare smaller sequence with larger sequence. Myers-Millers Algorithm is used to get global alignment.
- Local Alignment - This method does not assume to be two sequences are of same length because at the time alignment it considers the local region of sequence and not whole sequence. It only finds local regions with the highest level of similarity between two sequences without considering remaining part of sequences. The two sequences to be aligned can be of different lengths. Smith-Waterman algorithm is used to find out the local alignment.

II. Related Work

There are various techniques are available for retrieving score and alignment between biological sequences. The result of analysis used to discover functional, structural, evolutionary characteristic. The various biological sequence are very huge and for analysis of such huge sequences can be done by various algorithm like Smith-Waterman algorithm, Myers-millers algorithm, Fast LSA, CUDAlign 1.0, CUDAlign 2.0. To increase the efficiency of these algorithm there is use of GPU along with CPU in these algorithm.

T. F. Smith and M. S. Waterman [2] proposed Smith Waterman Algorithm, used to retrieves the optimal score and local alignment between two sequences. This algorithm totally based on Dynamic programming approach. In dynamic programming method applied to complex problem to solve that problem and get optimal solution by dividing the problem into small sub-problems then find the solution for each sub problem. Finally combine solution of all sub-problems to get overall solution. This algorithm is divides into three steps.

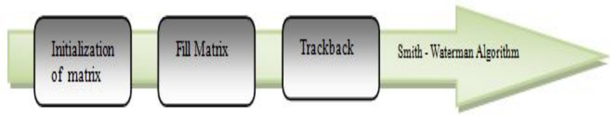

Fig 2: SW algorithm flow

First part of SW algorithm is to create the dynamic programming (DP) matrix and initialize it. Suppose the two sequences are seq1 = {A T C G } and seq2 = { T C G } of length m=4 and n=3 respectively then create DP of m+1 rows and n+1 column. Then initialize all cells the first row and column of matrix by 0 as shown in Figure 3.

		T	C	G
	0	0	0	0
A	0			
T	0			
C	0			
G	0			

Fig 3 : Initialization Step

Next part is to fill each cell of matrix and it can be done by using Equation (1). $H_{i,j}$ represents each cell of matrix. $S_{i,j}$ denote the similarity score according to match or mismatch for each instance and G is consider as Gap penalty.

$$H_{i,j} = \max \begin{Bmatrix} 0 \\ H_{i-1,j-1} + S_{i,j} \\ H_{i-1,j} + G \\ H_{i,j-1} + G \end{Bmatrix} \qquad (1)$$

Final stage of SW algorithm Trace back. It is used to find out optimal alignment which has maximum score. In trace back step, it starts from the highest score and continues until the minimum score as shown in Figure 4.

		T	C	G
	0	0	0	0
A	0	0	0	0
T	0	1	0	0
C	0	0	2	0
G	0	0	0	3

Fig 4: Trace back

Gotoh [3] modified the SW algorithm to include affine gap penalties. Because using a general form of gap penalty function slows down the algorithm, an affine gap penalty function is preferred. When using affine gap penalty function in dynamic programming, it only needs to differentiate between the case that the gap is first being introduced and the case that the gap is being extended. Hirschberg [4] present an algorithm which solve the

problem of quadratic time and space complexity of SW algorithm and proposed a linear space algorithm to compute the alignment of Longest Common Subsequence (LCS). Myers and Millers (MM)[5] algorithm that calculate the optimal global alignment in linear space. This algorithm of sequence alignment based on the Hirschberg's algorithm which uses divide and conquer procedure. The idea is to find the midpoint of Longest Common Sequence (LCS).

S. Aluru, N. Futamura, and K. Mehrotra [6] presented parallel algorithm which is based on prefix computations and handles pair wise comparison of biological sequences and able to handle both constant as well as affine gap penalty, full-sequence and subsequence matching of sequences. Fast LSA [7] is nothing but variant in MM algorithm that uses divide and conquer method. It performs parallelization in linear space.

III. Some GPU based Sequence Alignment Algorithm

Graphical Processing Unit(GPU) is one in which many cores are available that allows parallel execution of task. Several algorithms were design based on GPU helps to accelerate alignment of biological sequences.

Y. Liu, W. Huang, J. Johnson and S. Vaidya [8] present hardware implementation of double affine Smith Waterman (DASW) algorithm uses dynamic programming and implemented on a commodity graphics card. Other GPU based implementation [9],[10],[11] speed up alignment task but one of the restrictive characteristic of these algorithm that they cannot align two huge sequences.

CUDAlign 1.0 [12] algorithm able to handle large huge sequences to retrieve optimal score but not alignment. Main focus is on handling pair wise biological sequences. CUDAlign 2.0 is extended part of CUDAlign 1.0. This algorithm allows to calculate both score and alignment for huge sequences. CUDAlign 2.1 [1] includes optimization techniques to speed up the performance. Blockpurning is optimization step and goal of this step is to eliminate the calculation of cell that surely not belong to optimal alignment.

CUDAlign 2.1 follows algorithm introduced in CUDAlign 1.0 as base algorithm and then optimization techniques were introduced. As first part is to create DP matrix for two sequences as seq1 and seq2 of size m and n respectively. Matrix is divided into grid with n/rt *CB blocks where CB represent CUDA blocks , t is number of threads in each CUDA block and each thread is able to process r rows that means each block process rt rows. CUDAlign 2.1 divided into 6 stages-

1. Create DP matrix and obtain optimal score.
2. Partial traceback –to generate optimal alignment.
3. Splitting partition to create more crosspoints.
4. Apply MM algorithm with balance splitting
5. Concatenate all result of partition to obtain full alignment.
6. Representation of alignment.

IV. Conclusion

Sequence Analysis is key operation in Bioinformatics which is essential to discover functional, structural and evolutionary characteristic of newly generated sequence. Smith-waterman algorithm is used to obtain local alignment for biological sequence. Now a day's GPU receive lot of attention. For sequence analysis, many techniques introduce the use of GPU to make efficient implementation of Smith Waterman Algorithm. CUDAlign 2.1 is SW implementation based on GPU that is divided into six stages which introduce optimization techniques like block purning helps to speed up the performance.

Acknowledgment

The author wish to thank MET's Institute of Engineering Bhujbal Knowledge City Nasik, HOD of computer department, guide and parents for supporting and motivating for this work because without their blessing this was not possible.

References

[1] Edans Flavius de O. Sandes and Alba Cristina M.A. de Melo, Senior Member, IEEE, "Retrieving Smith-Waterman Alignments with Optimizations for Megabase Biological Sequences Using GPU," *IEEE transactions on parallel and distributed systems*, vol. 24, no. 5, may 2013.

[2] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences," *J. Molecular Biology*, vol. 147, no. 1, pp. 195-197, Mar. 1981

[3] O. Gotoh, "An Improved Algorithm for Matching Biological Sequences," *J. Molecular Biology*, vol. 162, no. 3, pp. 705-708, Dec. 1982.

[4] D.S. Hirschberg, "A Linear Space Algorithm for Computing Maximal Common Subsequences," *Comm. ACM*, vol. 18, no. 6, pp. 341-343, 1975.

[5] E.W. Myers and W. Miller, "Optimal Alignments in Linear Space," *Computer Applications in the Biosciences*, vol. 4, no. 1, pp. 11-17, 1988.

[6] S. Aluru, N. Futamura, and K. Mehrotra, "Parallel Biological Sequence Comparison Using Prefix Computations," *J. Parallel Distributed Computing*, vol. 63, no. 3, pp. 264-272, 2003.

[7] A. Driga, P. Lu, J. Schaeffer, D. Szafron, K. Charter, and I. Parsons, "FastLSA: A Fast, Linear-Space, Parallel and Sequential Algorithm for Sequence Alignment," *Algorithmica*, vol. 45, no. 3, pp. 337-375, 2006.

[8] Y. Liu, W. Huang, J. Johnson, and S. Vaidya, "GPU Accelerated Smith-Waterman," *Proc. Sixth Int'l Conf. Computational Science (ICCS)*, vol. 3994, pp. 188-195, 2006.

[9] W. Liu, B. Schmidt, G. Voss, A. Schroder, and W. Muller-Wittig, "Bio-Sequence Database Scanning on a GPU," *Proc. 20th Int'l Conf. Parallel and Distributed Processing (IPDPS)*, 2006.

[10] S. Manavski and G. Valle, "CUDA Compatible GPU Cards as Efficient Hardware Accelerators for Smith-Waterman Sequence Alignment," *BMC Bioinformatics, 9(Suppl 2)*, 2008.

[11] Y. Liu, D. Maskell, and B. Schmidt, "CUDASW++: Optimizing Smith-Waterman Sequence Database Searches for CUDA-Enabled Graphics Processing Units," *BMC Research Notes*, vol. 2, no. 1, p. 73, 2009.

[12] E.F. de, O. Sandes, and A.C.M.A. de Melo, "CUDAlign: Using GPU to Accelerate the Comparison of Megabase Genomic Sequences," *Proc. 15th ACM SIGPLAN Symp. Principles and Practice of Parallel Programming (PPoPP)*, pp. 137-146, 2010.